Basic Scaling

Stokman, Frans; Schuur, Wijbrandt van

Published in:
Quality & Quantity

DOI:
10.1007/BF00154792

Publication date:
1980

Quality and Quantity, 14 (1980) 5–30

BASIC SCALING

FRANS STOKMAN and WIJBRANDT van SCHUUR

Department of Sociology and Faculty of Social Sciences,
University of Groningen, The Netherlands

Representational Measurement

Introduction

Scaling is sometimes defined as the actual process of assigning numbers to objects. We have a set of objects, crimes for instance, ranging from "murder" to "rape" to "joyriding on a bicycle". We want to find out a) whether it is at all possible to order all crimes in an unambiguous way according to their "seriousness"; b) if so, in which order the crimes are put. So we compare all pairs of crimes, and find for instance, that "murder" is more serious than "joyriding on a bicycle", but "murder" and "rape" are equally serious. Such a set of objects, together with their different relations, is denoted an *empirical relational system*. We also have a set of numbers and different numerical relations between them (like "5 is larger than 4", or "3 is equal to 3"). This is denoted a *numerical relational system*. To each object from the set of objects one number from the set of numbers is assigned.

The problem in scaling can now be formulated as one of assigning numbers to the objects in such a way that the empirical relations between the objects are isomorphic with the corresponding relations between the numbers. If there is a dominance relation between two objects (like "crime i is judged to be more serious than crime j") this is represented by the number n_i being larger than the number n_j. And if there is a relation of indifference or equivalence between two objects, this is represented by the two corresponding numbers being the same. The possibility of ordering crimes unambiguously according to their seriousness implies the possibility of assigning increasingly large num-

bers to the crimes. These numbers can then be used as "scale values" on a variable "seriousness of crimes".

The same applies to relations between pairs of objects: if pair (x_i, x_j) is more similar than pair (x_p, x_q), then $|n_i - n_j| < |n_p - n_q|$. (For instance, murder and rape are more similar than joyriding on a bicycle and shoplifting.) Similarity between crimes may be judged on the basis of seriousness of the crimes. The numbers, assigned to each crime, can then again be interpreted as the scale values of the crime on the variable "perceived seriousness of crimes".

Relations between objects from two sets

A person solving an arithmetic exercise can be represented in this terminology by object x_i dominating object y_j. In such situations it makes sense to distinguish two sets of objects: a set of persons $x_i \in X$ and a set of arithmetic exercises $y_j \in Y$. In this example, the ordinary school situation, objects of set X are compared with objects of set Y. We want to find numbers to represent both persons and exercises separately such that if a person solves the exercise, he is represented by a higher number than the exercise. The numbers assigned to persons may be interpreted as scale values on an underlying variable like "arithmetic capability". The numbers assigned to the exercises may be interpreted as scale values on the same underlying variable, and indicate the difficulty of the exercises.

An example of relations between pairs of objects, where each pair contains an object from two different sets, is the preference of a person for different types of products. If John likes Dutch beer better than English beer, we may denote this as the difference between John's ideal beer (x_i) and Dutch beer (y_j) is smaller than the difference between John's ideal beer and English beer (y_k). Then the following relation must hold between the corresponding numbers in the numerical relational system: $|n_i - n_j| < |n_i - n_k|$. The numbers, assigned to actual and ideal beers, may be interpreted as scale values on the underlying variable, that was used to evaluate the different types of beer (perhaps sweetness or alcohol content).

Classification of scaling models

We have given above four types of relations between objects, based upon two distinctions: relations between ob-

TABLE 1

Scaling models according to the classification of Coombs

Number of sets of objects	Relations: Between objects	Relations: Between pairs of objects
1	paired comparison models of Thurstone and Bradley, Terry and Luce.	Distance model of Torgerson non-metric distance models.
2	Guttman scaling model Mokken scaling model Rasch scaling model	(Multidimensional) preference models (e.g. unfolding)

jects and relations between pairs of objects; relations between objects from the same set of objects and relations between objects from different sets of objects. These four types of relations give rise to four different types of scaling models. This classification of scaling models stems from Coombs (1964) and has been very influential. Table 1 classifies some well-known scaling models according to this classification. For each of the models references and basic texts are given in the list at the end of the chapter.

Representation problems

It is not always easy to assign numbers to objects in such a way that the relation between the numbers is isomorphic with the relation between the objects. If crime i is judged to be more serious than crime j, crime j more serious than crime k and crime k in turn more serious than crime i, we are in trouble. Basically, we can approach this problem in two different ways:

1) find numbers corresponding to crimes i, j and k in such a way, that not too much damage is done in terms of inconsistency (for instance, give them all the same number);
2) conclude that there is no single dimension called "seriousness of crimes" (that is, conclude that the scaling model used is not appropriate).

As Coombs, Dawes and Tversky (1971, p. 32) comment:

> When a scaling model is applied as a *technique*, it assumes some particular measurement model. Departures from the model are regarded as random fluctuations or observed errors, and the purpose of scaling is to find numbers that provide the "best" fit between the model and the data in the presence of some error. The concept of best fit is usually defined in terms of some error function to be minimized, *e.g.* the sum of squared deviations in the method of least squares.
>
> When a scaling model is applied as a *criterion*, it is employed as a method of testing the descriptive validity of some measurement model. The purpose of scaling in this case is to discover whether the data can be fitted by the model. Consequently, when a scaling procedure is used as a technique, it tends to be insensitive to error as it is designed to yield numerical scale values regardless of whether the measurement model is satisfied or not. In contrast, when a scaling procedure is used as a criterion, it tends to be sensitive to error because its main objective is to detect departures from the theory. A scaling model can, clearly, be used as a technique and as a criterion.

Measurement as a gift of the valid model

A scaling procedure as technique requires estimation of the parameters of a model. There are two different types of model, those which take measurement error into account, and those which don't. The first type of model is called "probabilistic" or "stochastic", and the second type is called "deterministic". We shall give an illustration of the difference between these two types of models later. A scaling model as criterion requires derivation of restrictions on the data from the model that can be used for the definition of one or more goodness-of-fit criteria. A scaling model is said to be valid as a criterion if the fit between the data and the model is better than some preconceived value.

Goodness-of-fit calculations are based on the numbers that provide this fit (*i.e.*, estimates of the parameters of the model). Hence, given a good fit between data and model, we thereby also have the numbers to represent the objects. The numbers are a gift of the valid model, so to speak. Scaling models in this sense can be regarded as

miniature theories. Transitivity of seriousness of crimes, for instance, is by no means trivial. We can regard it as a theoretical postulate to be tested by a goodness-of-fit criterion.

Uni- and multidimensional scaling

In the scaling models discussed so far, we have implicitly assumed that only one number is assigned to each object. This need not be the case. Often relations between pairs of objects may need more than one number per object for a good-fitting representation. In many cases the assignment of numbers to objects can best be seen by a geometrical analogy. When each object gets one number, we can represent the objects along a line, a one-dimensional continuum. When each object gets two (or more) numbers, we have to represent the objects as points in a two (or more) dimensional space, where the numbers assigned to each object can be seen as the values on the coordinate axes of the multidimensional space.

For instance, the relations between three objects, A, B and C lead to the following order of similarity between the pairs of objects: AC : AB : BC (A and C most similar). If we assume an inverse relationship between similarity and distance, the objects A and C must be represented closest to each other on the line, and the objects B and C must be farthest apart. The three objects can be represented (e.g.) by the numbers A (8), B (17) and C (0). So only one number is needed for each object. Given a fourth object D, and the following order of similarities: AD : AC : AB : CD : BD : BC (A and D most similar), it is no longer possible to represent the objects by one number. A perfect representation may be acquired, however, if we assign two numbers to each object: A (8, 0), B (17, 0), C (0, 0), D (8, 6).

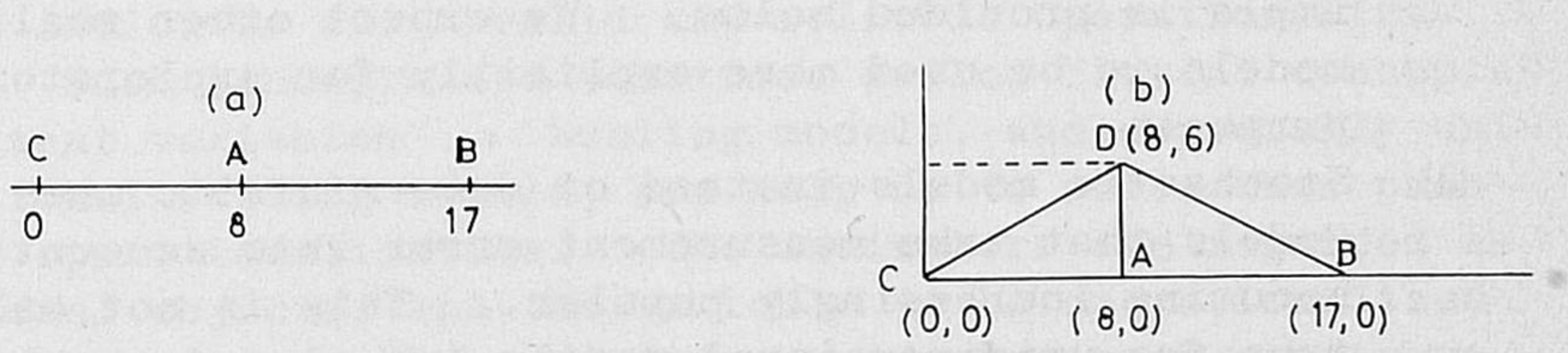


Fig. 1. (a) Three points in one dimension and (b) four points in two dimensions.

Scaling procedures that assign more than one number to each object are called multidimensional scaling procedures.

Admissible transformations and uniqueness

Apart from the representation problem we have the uniqueness problem. Given a particular measurement procedure, how much freedom do we have in assigning numbers to objects? If the only relations observed between objects are dominance relations (like the "seriousness of crimes" example) any monotone transformation of the numbers keeps the representation of the dominance relation intact. The scale type is said to be ordinal. If dominance *or* equivalence relations exist between pairs of objects, an admissible transformation of the numbers assigned to the objects must preserve the order of the difference between the scale values. All interval scales, where any positive linear transformation is allowed, have this property.

Prospects

Scaling models and procedures are still being developed. From our (subjective) point of view, we see three different areas of emphasis in the near future:

1) Scaling procedures for detecting structure. It need not always be the case that the latent concepts we can operationalize were already theoretically specified. Instead of starting from a theoretical concept, we may start from a number of indicators and find out what their structure is. It may turn out that a number of indicators (not necessarily all of them) can be represented as a good fitting uni- or multidimensional scale. These exploratory purposes of scaling are most emphasized in the stochastic unidimensional scaling model of Mokken, of which an example is provided below. We expect other scaling models to be used more explicitly for exploratory purposes.
2) Stochastic models instead of deterministic ones. Models that take measurement error into account are becoming increasingly popular. This is not only true for unidimensional models but also for multidimensional models of similarity and preference analysis.
3) Models based on empirical relations other than domi-

nance and indifference. Dominance or indifference relations are not the only possible relations between objects or pairs of objects. Empirical relations that are represented by addition, multiplication or combinations of both can be postulated. Certain "conjoint measurement" models are based on these relations. Recently, more emphasis has been placed on hierarchical relations between objects, that may be represented by additive trees (Sattah and Tversky, 1977) or hierarchical trees (Carroll and Pruzansky, 1975).

Scaling as Attitude Measurement

It is not possible to give a full account of every scaling model in this single chapter. Overviews exist for the more classical models, including Torgerson (1958), Coombs (1964), Coombs, Dawes and Tversky (1971), Dawes (1972), Scheuch and Zehnpfennig (1974) and the Basic Scaling monograph of the ECPR Summer School. New developments are highly scattered, but a number of them can be found in *Psychometrika*. In the remainder of this chapter we elaborate one model, based on the dominance relation between objects from different sets. This model has numerous applications, especially in the field of attitude measurement, where the two sets of objects are persons and statements which operationalize an attitude.

Introduction

Social scientific theories often contain concepts or variables that are difficult to operationalize, like status, intelligence, alienation or authoritarianism. The difficulty in measuring these concepts is that we cannot find one single indicator for a valid and reliable operationalization. Such concepts are sometimes called "latent concepts" or "latent variables". Scaling models, and especially unidimensional scaling models, can help us to solve this problem of operationalization. The indicators are regarded as individual objects, and the variable to be operationalized can be regarded as the underlying unidimensional continuum. Attitude measurement is a good case in point, but by no means the only one: the attitude "political efficacy" in political science literature, is often operationalized by a combination of responses to such questions as:

"So many other people vote in the national elections that it does not matter much to me whether I vote or not".
"Because I know so little about politics, I should not really vote".
"Sometimes politics and government seem so complicated, that a person like me cannot really understand what is going on".
"Members of Parliament do not care much about the opinion of people like me".

A person who disagrees with any of these statements is said to give the "efficacious" response. The number of statements a person disagrees with, ranging from zero to four, is used as his value on the variable "political efficacy". It is called his "scale score". The combination of these four statements is called a "scale". The scaling model we use in this case relates objects (not pairs of objects) of two different sets: a person is compared with an attitude statement. If the person gives the "efficacious" response he is said to dominate the statement. If he does not give the efficacious response, the statement is said to dominate him. This model is elaborated below.

Index measurement:
equal appearing intervals and the Likert scale

In earlier scaling procedures, the assignment of numbers to objects has not always been related to empirically observed dominance or equivalence relations between subjects and attitude statements. We shall give two examples of this: the method of equal appearing intervals, developed by Thurstone in 1928 and the Likert scale, developed by Likert in 1932. In the method of equal appearing intervals, judges are asked to select a number of statements to operationalize an attitude as an underlying latent variable. Each judge sorts every statement on an eleven point scale, ranging from extremely unfavorable to extremely favorable. Those (about twenty) statements are selected that are distributed equally along the attitude continuum according to the most unambiguous judgements. The standard deviation of the judgement distribution may be used as a measure of ambiguity. A subject's attitude score is the mean (according to Torgerson) or median (according to Edwards) value of the statements the subject endorses.

In the Likert scale procedure statements (also called "items") are used with five response categories (strongly

agree, agree, uncertain, disagree and strongly disagree). These response categories are coded as 1, 2, 3, 4, 5 for statements phrased in the direction of the concept (e.g., the "efficacious" response), and are coded as 5, 4, 3, 2, 1 for statements phrased in the opposite direction. Those statements are selected that have a sufficiently high item-test correlation. The scale score of the respondent (also called the "test") consists of the sum of the values of the response categories he has used. The main difference between these two procedures is that in the first procedure numbers are assigned to statements, whereas in the second numbers are assigned to response categories. Note that in both cases the assignment of numbers to objects is not connected with any empirically-observed dominance or indifference relation between the objects. These scaling procedures, where numbers are assigned to subjects and statements on a more or less arbitrary basis, are called index measurement procedures. Scales formed on the basis of index measurement are also called rating scales. Scaling procedures where the assignment of numbers to objects is related to empirically-observed relations are called representational measurement procedures. Dawes (1972) makes the distinction between index measurement and representational measurement the basic distinction of his overview of attitude measurement techniques. Whereas future observations of empirical relations can be predicted from the application of representational measurement models, this is not the case with index measurement. Instead, the justification of rating scales lies only in their utility.

The stochastic cumulative model

We assume the existence of a *one-dimensional latent attribute* (e.g., an attitude, such as political efficacy) which can be represented as a line or underlying continuum, as in Figure 2. We assume that each *subject* has a certain, unknown *value* on this continuum, giving the amount of the attribute that a subject possesses. We denote this value by the Greek letter θ, as in Figure 3.

We assume a set of *items* that are related to the attribute, that is, are *homogeneous* with respect to it. These items have *two response alternatives*: one alternative that expresses the attitude (positive or "+" alternative) and one alternative that does not express the attitude (negative or "-" alternative). Items with more response alternatives

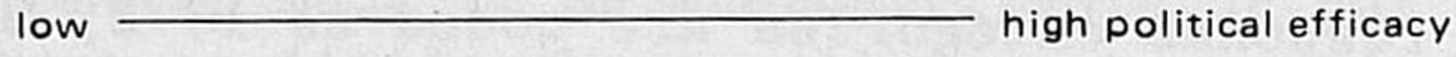


Fig. 2. One-dimensional latent attribute.

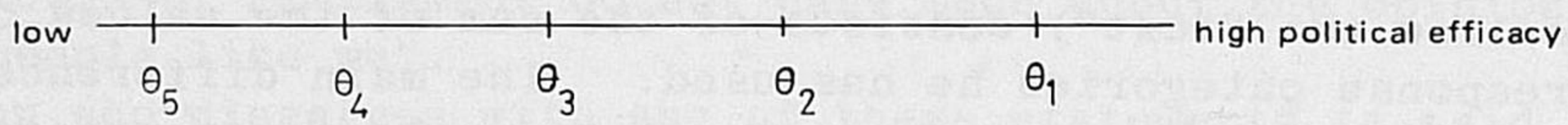


Fig. 3. Subject values on continuum.

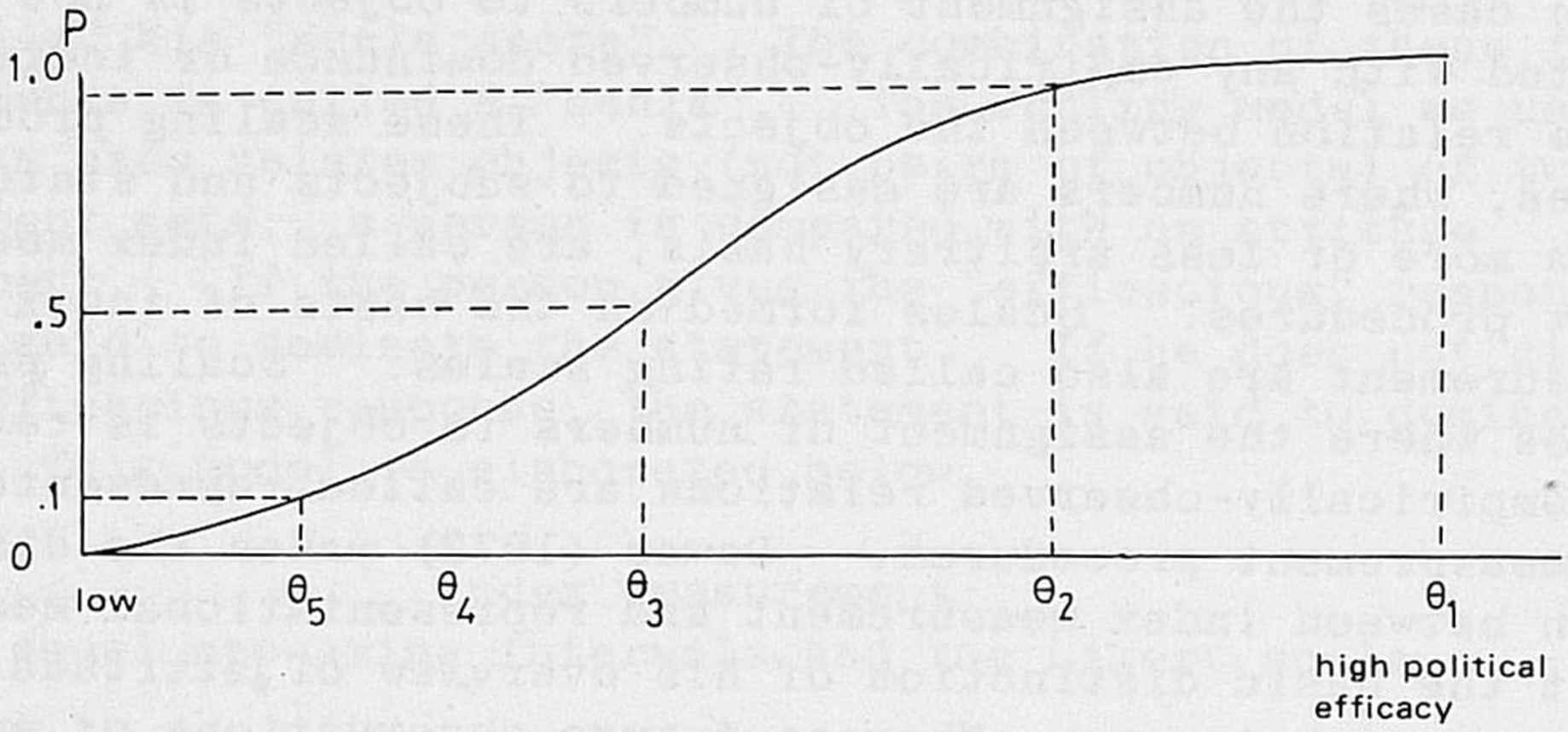


Fig. 4. Trace line of an item.

are dichotomized. We assume for each item that the probability of a positive response increases monotonically with the (unknown) value θ. Subjects with a higher value of θ will be more likely to give the positive response than subjects with a lower value of θ. In Figure 4, the probability of a positive response is given on the vertical axis and the underlying continuum is given on the horizontal axis. We see that the probability of a positive response increases monotonically or at least does not decrease. For subject 5 with value θ_5 this probability p is 0.10, for subject 3 with value θ_3 p = 0.50 and for subject 1 with value θ_1, p is almost 1. The function that for an item gives the probability of a positive response for the different values of θ is denoted a *trace line*.

In Figure 4, only one trace line is given. However, each item has such a trace line. We assume that certain items are more *difficult* than others, i.e., that to obtain a positive response, certain items require a larger amount of

the latent attribute. The probability of a positive response for such a more difficult item is smaller than (or at the most equal to) that of an easier item. Thus we get Figure 5 with trace lines for four items.

In Figure 5 an (unknown) value on the latent variable is given for each item. Each item gets the value corresponding to the value of the subject giving the positive response with a probability of .50. This value is denoted δ. The items are ordered from difficult (δ_1) to easy (δ_4). The value of item 3, δ_3, is equal to that of subject 3, because this subject gives the positive response to item 3 with a probability of .50 ($\delta_3 = \theta_3$). It is required not only that the trace lines do not decrease over the continuum, but also that they do not intersect. For each θ value (e.g., θ_3) it is required that the probability of positive response on item 1 $\leqslant$ probability of positive response on

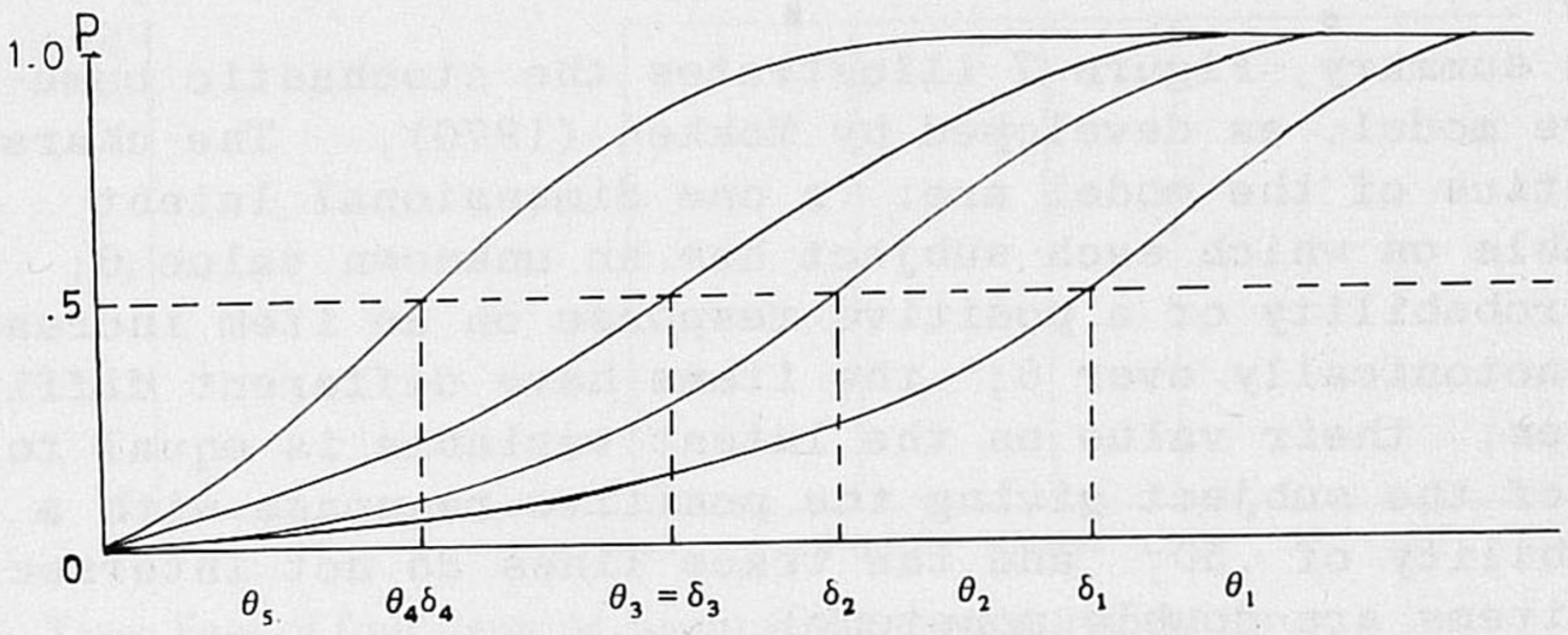


Fig. 5. Trace lines for a set of four items.

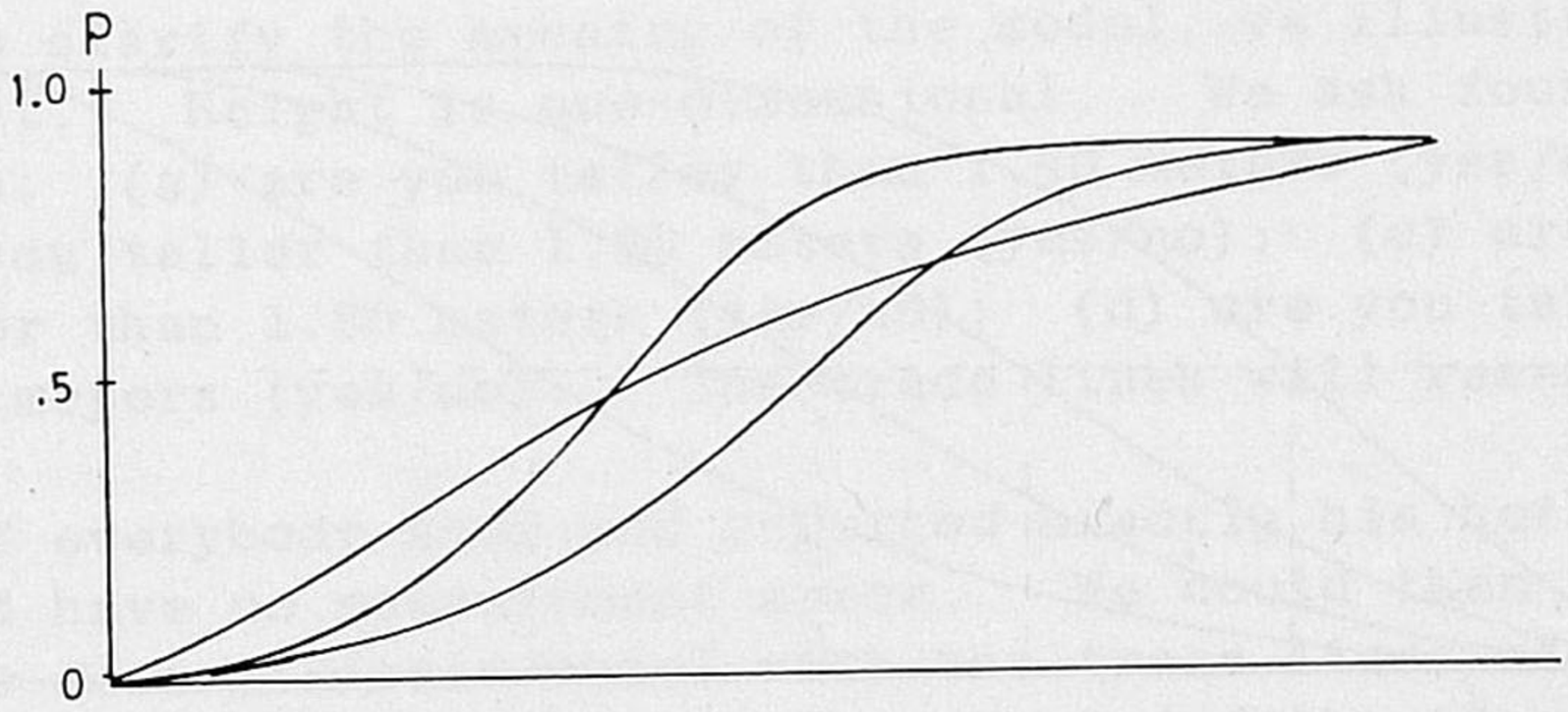


Fig. 6. Intersecting trace lines (not allowed).

item 2 ≤ probability of positive response on item 3, and so on, where item 1 is more difficult than item 2, and so on. We therefore have a *double monotonicity*: one of each trace line over the θ-values and one of the probabilities of a positive response for each θ-value. If the trace lines intersect, as in Figure 6, we cannot say which item is more difficult. For certain θ-values one of the items is more difficult (has a smaller probability of a positive response) whereas for other θ-values the other item is more difficult.

Here, as well as throughout the whole stochastic cumulative scaling model, we assume a subject's responses to different items are statistically independent. This means that the probability of responding positively to a pair of items for any subject is the product of his probabilities responding positively to each item separately. This assumption, also crucial in classical test theory, is called the assumption of "local stochastic independence".

In summary, Figure 7 illustrates the stochastic cumulative model, as developed by Mokken (1970). The characteristics of the model are: a one dimensional latent variable on which each subject has an unknown value θ; the probability of a positive response on an item increases monotonically over θ; the items have different difficulties; their value on the latent variable is equal to that of the subject giving the positive response with a probability of .50; and the trace lines do not intersect (the items are double monotone).

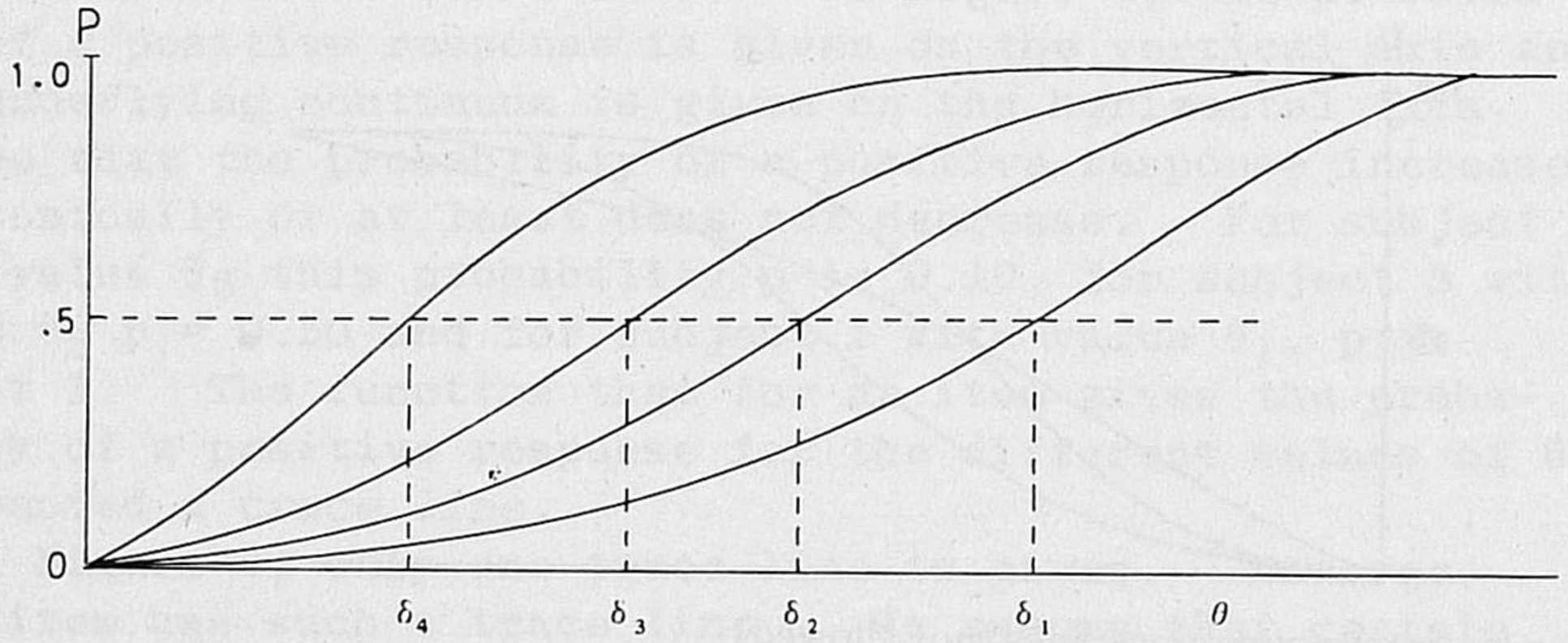


Fig. 7. The stochastic cumulative scaling model.

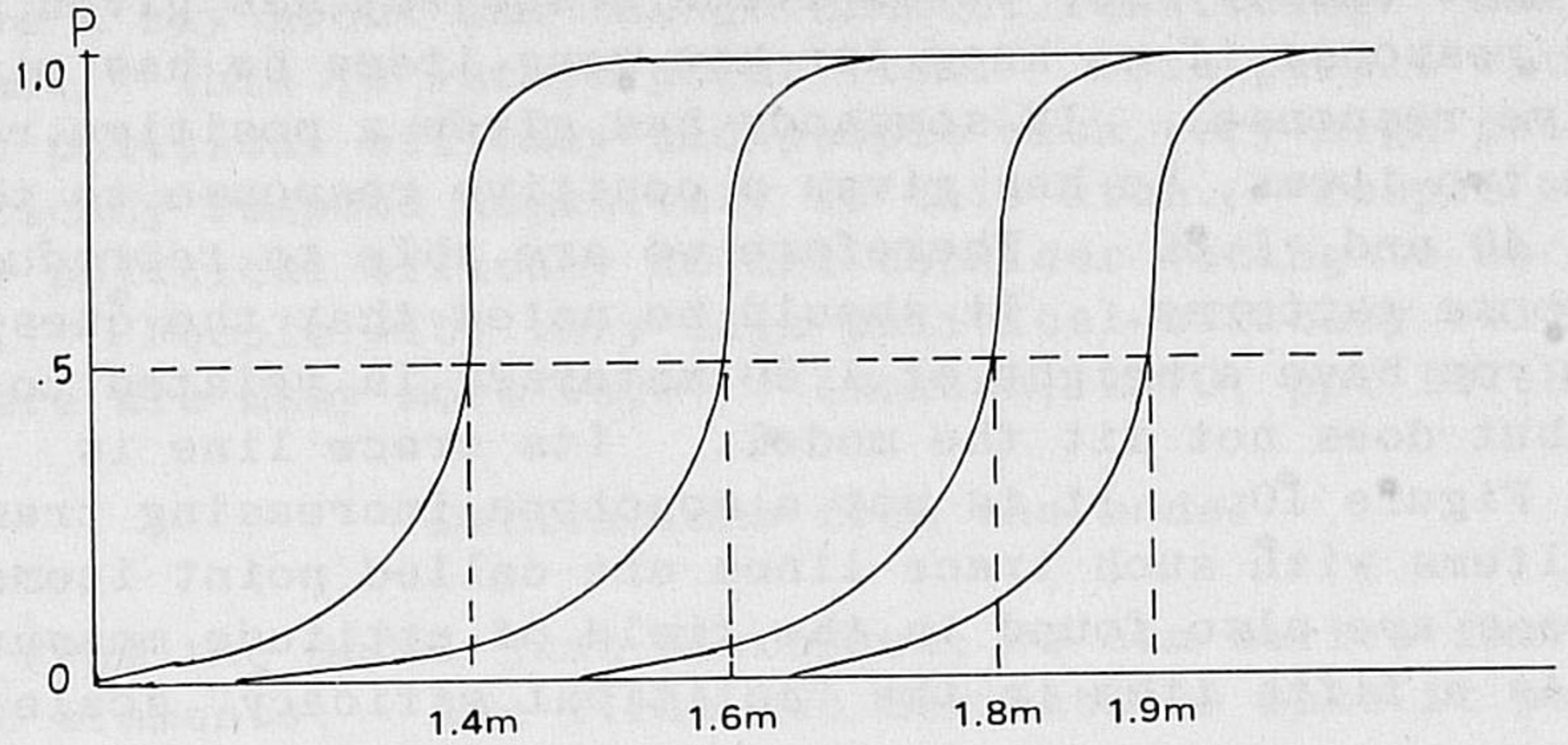


Fig. 8. Trace lines of four items on height.

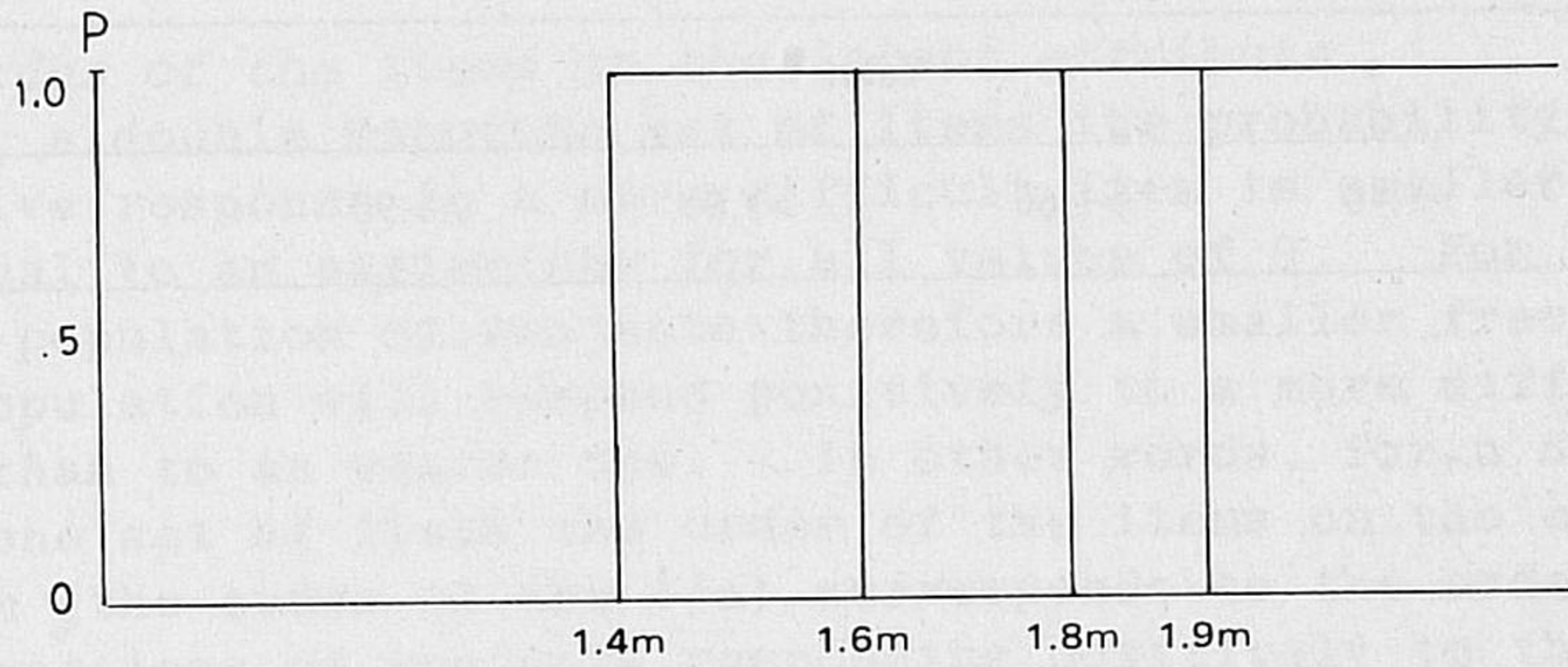


Fig. 9. Trace lines of four items on height without measurement error.

To clarify the meaning of the model, we illustrate it for height. Height is one-dimensional. We ask four questions: (a) are you taller than 1.40 meters (yes/no); (b) are you taller than 1.60 meters (yes/no); (c) are you taller than 1.80 meters (yes/no); (d) are you taller than 1.90 meters (yes/no). The trace lines will resemble Figure 8.

If everybody *knew and reported exactly* his height, we would have no measurement error. We could then apply Guttman's *deterministic* model with the trace lines of Figure 9.

In this deterministic case only 5 (of the $2^4 = 16$) response patterns can appear (in general with k items we have $k + 1$ response patterns out of the 2^k possible ones).

These patterns are given in Table 2. Moreover, in this case we know exactly for which items a subject has given the positive response if we know for how many items he has given a positive response. If somebody has given a positive response on two items, he has given a positive response to the items >1.40 and >1.60. Therefore we are able to *reproduce the response patterns*. It should be noted that the question "Do you have a height of 1.80 meters?" is related to height, but does not fit the model. Its trace line is given in Figure 10. It is not a monotone increasing trace line. Items with such trace lines are called point items. Point items are also found in the field of attitude measurement. As a fifth item in the "political efficacy" scale,

TABLE 2

Response patterns if no measurement error existed

	Value			
Subject	>1.40	>1.60	>1.80	>1.90
1	–	–	–	–
2	+	–	–	–
3	+	+	–	–
4	+	+	+	–
5	+	+	+	+

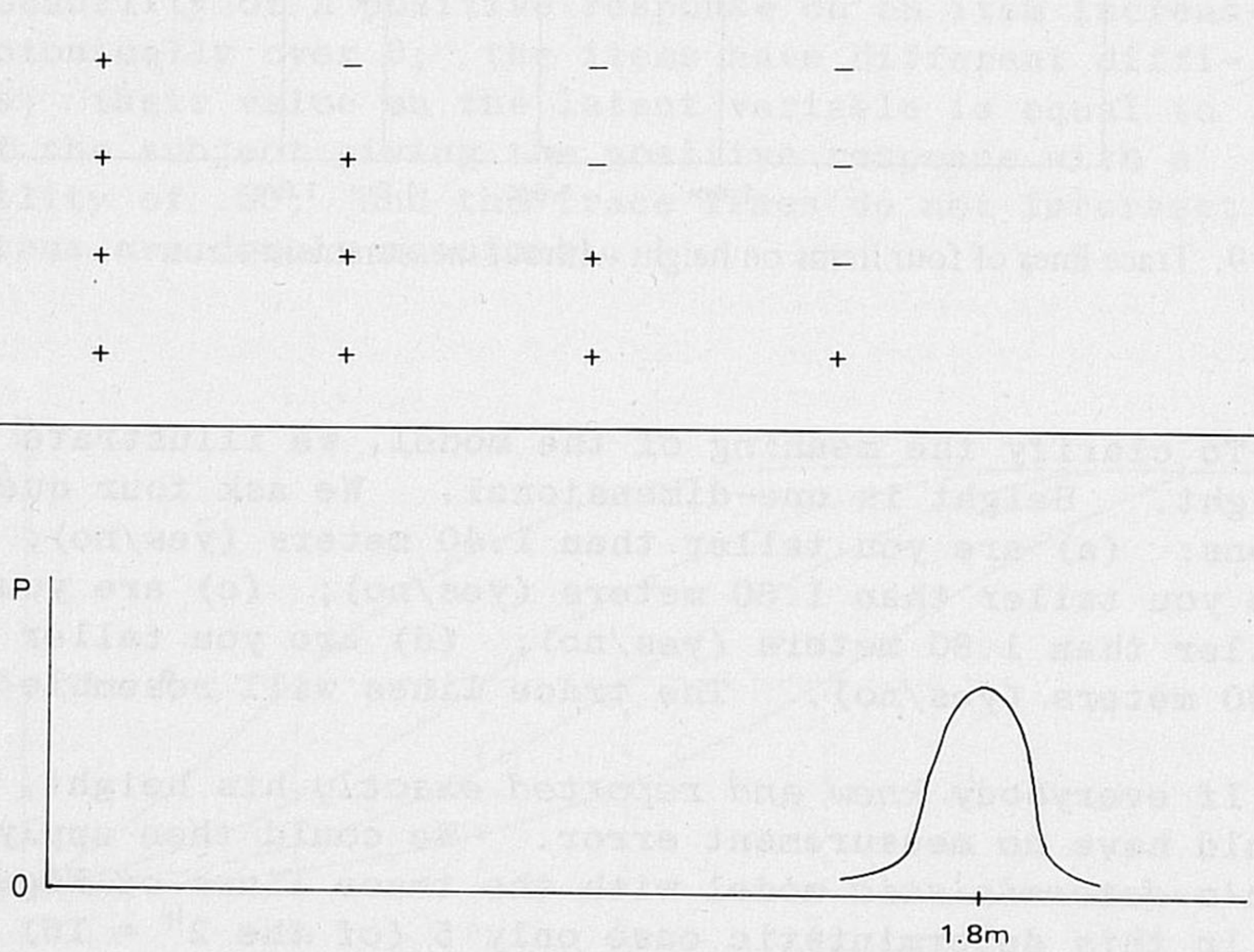


Fig. 10. Trace line of the question 'Do you have a height of 1.80 meter?'

the item "Voting is the only way that people like me can have a say about how the government runs things" has been used. This is such a point item. Both people with very low political efficacy and people with very high political efficacy respond negatively to this item. People with very low political efficacy do not consider voting to be a way at all. People with very high political efficacy believe that there are many more ways. (Mokken, 1970, pp. 137-8.)

Derivations from the model

From the model just discussed, we can derive four different elements. The first two derivations make it possible to determine the order of the items and the subjects on the latent attribute (i.e., to get ordinal *measures* of items and subjects). The other two derivations give restrictions for the possible responses and can therefore be used to *test* the model.

The order of the items on the latent attribute

For a double monotone set of items the probability of a positive response to a more difficult item is smaller than or equal to an easier one for all values of θ. For the total population of subjects therefore a smaller fraction of the population will respond positively to a more difficult item than to an easier one. In other words, for a double monotone set of items the order of the items on the continuum (the order of the δ's) corresponds to the order of the fractions of subjects responding positively to the items. In general, this fraction in the population (denoted π_i in which i denotes the i-th item) is unknown. It can be estimated from the sample fraction p_i, the proportion of respondents in the sample that respond positively to item i. Thus, for a double monotone set of items we have

$$\delta_i > \delta_j <=> \pi_i < \pi_j \qquad (1)$$

The order of the subjects on the latent attribute

Because for each item the probability of a positive answer increases with the value θ, the order of the subjects on the continuum can be estimated from the number of positive responses he gives. This is denoted the *summation score*. With k items, it gives a *partial order* of the subjects in k + 1 categories.

TABLE 3

Cross-tabulation of two items

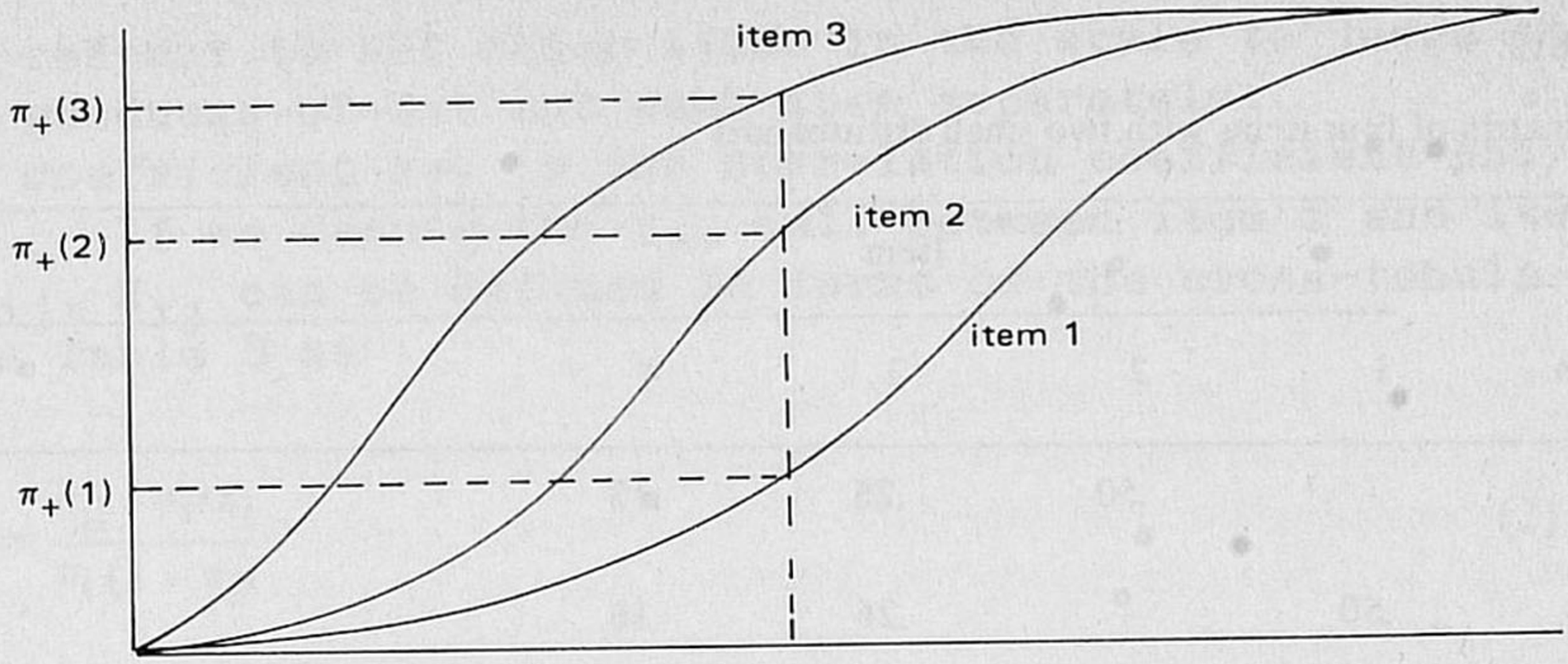
	Item 2		
Item 1	+	−	
+	π_{++}	π_{+-}	π_1
−	π_{-+}	π_{--}	$1-\pi_1$
	π_2	$1-\pi_2$	1

Positive correlation between each pair of items

The responses to each pair of items with monotonically increasing trace lines are positively correlated. Let us look at the cross-tabulation of two items, using the symbol π for the unknown population proportion in each cell, as in Table 3. The probability of a positive response to item 1 is π_1; the probability of a positive response to item 2 is π_2. Because item 1 is more difficult than item 2, $\pi_1 \leq \pi_2$. The probability of positive responses to both item 1 and item 2 is π_{++}. If the responses to items 1 and 2 are independent, the probability of two positive responses is $\pi_1 \cdot \pi_2$. Positive correlation implies that $\pi_{++} > \pi_1 \cdot \pi_2$. Again the unknown π's can be estimated on the basis of the proportions in the sample. Maximum correlation exists if $\pi_{+-} = 0$ i.e., if all subjects giving the positive response to the more difficult item also give the positive response to the easier item. The lower the fraction π_{+-}, the stronger the correlation. The π_{+-} cell is therefore sometimes called the error cell.

Monotonicity in the P-matrix and P_o-matrix

Above we have seen that positive correlation between two items exists if the trace lines increase monotonically. It does not imply, however, that they may not intersect. For a *double* monotone set of items, a necessary, but not sufficient, condition can be formulated. This can best be illustrated by an example of three items with different difficulties and non-intersecting trace lines (see Figure 11). For a given subject, the probability of responding positively to both item 3 ($\pi_+(3)$) and item 2 ($\pi_+(2)$) is higher than



For each θ: $(\pi_+(3)\cdot\pi_+(2)) > (\pi_+(3)\cdot\pi_+(1)) > (\pi_+(2)\cdot\pi_+(1))$ hence for the group of respondents as a whole $\pi_{++}(3,2) > \pi_{++}(3,1) > \pi_{++}(2,1)$.

Fig. 11. Three double monotone trace lines.

the probability of responding positively to both item 3 and item 1 ($\pi_+(1)$), which is in turn higher than the probability of responding positively to both item 2 and item 1. If the trace lines do not intersect, this relation between the three probabilities holds for every subject. The positive response to both easiest items should be highest and should decrease for item pairs with increasing difficulty.

If we order the items from difficult to easy and insert in a matrix for each pair of items i,j their π_{++} (i, j) proportions (see Table 3), then these π_{++}(i, j)'s should increase from left to right and from top to bottom. Such a matrix is denoted the *P-matrix*. Note that the order relation of the π_{++} (i, j)'s in the P-matrix is a necessary but not sufficient condition for double monotonicity. Estimat-

TABLE 4

P-matrix of four items

Item	1	2	3	4
1	–	.27	.30	.32
2	.27	–	.36	.38
3	.30	.36	–	.64
4	.32	.38	.64	–
Item difficulty (p_i)	.36	.41	.69	.82

TABLE 5

P_0-matrix of four items with two small disturbances

	Item			
Item	1	2	3	4
1	–	.50	.25	.13
2	.50	–	.26	.16
3	.25	.26	–	.12
4	.13	.16	.12	–
Item difficulty (p_i)	.36	.41	.69	.82

ing the unknown π's on the basis of the sample fractions, we might obtain a P-matrix like that in Table 4. Small departures from the double monotone requirement in the P-matrix may be due to sampling error.

For a double monotone set of items a similar necessary (but not sufficient) condition can be formulated for the P-null matrix (P_0-*matrix*), in which all π_{--} are inserted. If we order the items from difficult to easy and insert in the matrix for each two items their π_{--} fraction (see Table 3), then these π_{--}'s should decrease from left to right and hence from top to bottom. On the basis of sample estimates we might obtain a P_0-matrix like Table 5, which contains two small disturbances.

Scalability criteria: goodness of fit

The two last derivations have been used to develop a number of scalability criteria to test the goodness-of-fit of the model. In particular, the positive correlation between each pair of items has been used as a basis for different coefficients of scalability, including

1) the coefficient H_{ij} for each two items (i and j) in the scale;
2) the coefficient H for the scale as a whole;
3) the coefficient H_i for each item in the scale with

respect to all other items in the scale to judge the goodness-of-fit for each item separately.

The *coefficient* H_{ij} is the correlation coefficient phi/phi_{max}. If we denote the π_{++} cell between item i and item j as π_{ij}, H_{ij} can be defined in terms of the cross-tabulation in Table 3 as

$$H_{ij} = \frac{\pi_{ij} - \pi_i \cdot \pi_j}{\pi_i (1 - \pi_j)} \qquad (2)$$

H_{ij} is 1 in case of maximal correlation ($\pi_{+-} = 0$); H_{ij} is 0 in the case of statistical independence of the responses to the two items. In actual practice H_{ij} will be estimated ($\hat{H}_{ij}$) using sample proportions.

With the H_{ij}-coefficients between each pair of items we have a measure of positive correlation between those items, but not yet a clear criterion for all items together, that is, for the scale as a whole. For this purpose *Loevinger's coefficient of homogeneity* H is used, which is a weighted average of all H_{ij}'s. On the basis of the cross-tabulation in Table 3 H can be defined as follows:

$$H = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (\pi_{ij} - \pi_i \cdot \pi_j)}{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \pi_i(1-\pi_j)} \qquad (3)$$

H is also estimated as $\hat{H}$ using the sample proportions.

To judge the scalability of a set of items it is important to know whether *all* items fit the scale or not. It might well be that a low scalability of a set of items (a low H coefficient) is due to the fact that one item does not fit. For this reason the *item coefficient* H_i is introduced. It is a weighted average of the H_{ij}'s of that item with all other items in the scale:

$$H_i = \frac{\sum_{j=1}^{k} (\pi_{ij} - \pi_i \cdot \pi_j)}{\sum_{j=1}^{i-1} \pi_j(1-\pi_i) + \sum_{j=i+1}^{k} \pi_i(1-\pi_j)} \qquad (4)$$

As before, H_i is estimated ($\hat{H}_i$) using the sample proportions.

On the basis of these coefficients the following *definition of a scale* is given:

A set of items constitutes a scale if:

1) All $\hat{H}_{ij}$'s > 0
2) All $\hat{H}_i$'s $\geq$ c (and therefore $\hat{H} \geq$ c) where c is a positive constant. In actual practice the following criteria may be used:

$\hat{H} \geq .50$ a strong scale
$.40 \leq \hat{H} < .50$ a medium scale
$.30 \leq \hat{H} < .40$ a weak scale
$\hat{H} < .30$ no scale

In addition to this two other conditions are formulated:

3) $\hat{H}$ and all $\hat{H}_i$'s should be *significantly greater* than 0 at a level of confidence chosen by the researcher. Given the marginals of the items, $\hat{H}$ and the $\hat{H}_i$'s are asymptotically normally distributed for large numbers of subjects under the null hypothesis of random response ($H = H_i = 0$).
4) The trace lines should be *double monotone*. As a check the P-matrix and P_0-matrix should be inspected for monotonicity. Unfortunately no clear criteria have as yet been developed to determine whether small disturbances in monotonicity (because of sampling errors) are acceptable or not.

As criteria of goodness-of-fit the coefficients are helpful because they enable us to judge the scale as a whole (H) as well as the scalability of each item separately (H_i). Moreover, because of the known distributions of $\hat{H}$ and $\hat{H}_i$ under the null hypothesis of random response we are able to test whether the sample values are significantly larger than 0 and therefore are unlikely to be due to random fluctuations.

Thus, using the coefficients H_{ij}, H_i and H, based on the positive correlation between the items, and the P-matrix and P_0-matrix as a check of the double monotonicity, we are able to *test* the goodness-of-fit of the model. If these tests are satisfactory (we need not reject our theory) we can use the proportions p_i as a measure to order the items on the latent attribute and the summation score as a measure to order the subjects on the latent attribute. If the tests are not satisfactory, we have to reject our theory and at the same time we have no measures to order the items and subjects. The measures are therefore directly related to and derived from our theory about the construct.

Robustness

If we have a satisfactory cumulative scale of items on the basis of the above criteria, we may want to investigate whether the structure of the scale is approximately the same across different populations or for different subgroups in a population. The *robustness* of a scale over a number of subgroups in a population (or over different populations) can be investigated by testing the equality of H (or indeed the H_i) for different subgroups. A statistic T has been introduced, which has approximately a χ^2 distribution (with p-1 degrees of freedom if p populations (subgroups) are compared) under the null hypothesis that the H or H_i coefficients are equal in all populations. With the statistic T we therefore can test whether the fluctuations of the $\hat{H}$ or $\hat{H}_i$ coefficients in different samples can be treated as random or not.

Item selection

Until now, we have considered only the evaluation of a set of items as one scale. The H coefficient enables us to evaluate the scale as a whole while the H_i coefficients reflect the goodness-of-fit of the items separately. In exploratory research we may want to search for scales which can be constructed from a pool of items. For that reason a multiple (stepwise) scaling procedure has been devised, in which items are successively combined in a stochastic cumulative scale, as long as the conditions for such a scale ($H_{ij} > 0$, $H_i \geq c$) are fulfilled. For the remaining items that did not form part of this first scale, the procedure starts again, searching for items to form a second stochastic cumulative scale. Sometimes three or even more scales are found in this way. The multiple scaling procedure as well as the introduction of the item coefficients H_i reflect the idea that *item selection* is an essential element of scaling.

Analyses

Political efficacy was previously mentioned as an example to which the model can be applied. In survey research the model has successfully been applied to other political and social attitudes, political knowledge and political participation. It should be realized, however, that the model can also be applied in quite different contexts. One such application is the elaboration of a *cumulative leadership*

model in the General Assembly of the United Nations on the basis of sponsorship of resolutions by delegations. It should be noted that the items in the model become delegations in this model, whereas the subjects in the model are now proposals. The active and passive sets of the scaling model are therefore reversed in this application. A positive response to an item corresponds to the fact that the proposal is (co-)sponsored by a delegation.

The cumulative leadership model is based on the notion of the existence of one hierarchy of leaders within a group. We cannot speak of one hierarchy if there is a division of labor in the group, i.e., if one subgroup of delegations is active regarding one set of group goals and other subgroups are concerned with other sets. In such a situation we shall probably discover several hierarchies of leadership, depending on the kind of group goals we consider. This can be thought of as issue-specific leadership within the group. One hierarchy of leadership within a group implies active participation of certain delegations regarding all group goals, i.e., general leadership. In the case of our sponsorship data we might speak of one hierarchy of leadership in a group if sponsorship is cumulative. "Leaders" in the group are then active in a broad field. They sponsor a large number of proposals over the whole range of group-goal-related issues. "Followers" in the group are more reluctant to sponsor proposals. They only become involved if proposals are more highly salient to group goals. The more salient a proposal is to the group goals, the more delegations will co-sponsor it. If delegations lower in the hierarchy sponsor a proposal, the leaders of the group are also very likely to be on the list of sponsors. For our sponsorship data we may thus specify the following conditions for the existence of one hierarchy of leadership in a group:

1) The different proposals can be ordered along a single continuum, representing "saliency to group goals". We expect that each proposal has a given, but unknown, value on that "underlying" variable. Let that unknown value be θ.
2) A delegation is more likely to sponsor a proposal that is salient to its group goals than one that is marginally related to its group goals. The probability that a delegation will sponsor a proposal increases monotonically with the value θ, the unknown value of a proposal on the variable "saliency to group goals".

3) Each delegation also has an unknown value δ_i on that θ-axis. That value represents its reluctance to sponsor a proposal. The higher the delegation in the leadership hierarchy, the lower is its reluctance to sponsor, thus the lower is its value δ_i. For theoretical reasons we make the value δ_i equal to the value θ of the proposal that the delegation sponsors with probability of 0.5.
4) A proposal sponsored by a more reluctant delegation (a delegation lower in the hierarchy), is very likely to be sponsored also by a less reluctant delegation, (a delegation higher in the hierarchy).

The model can then be represented as in Figure 7, using four delegations with difficulties δ_1 to δ_4. This model has been used to investigate leadership and group formations among developing nations within the United Nations over the period 1950-68 (Stokman, 1977). Table 6 contains, as an example, the results for the period 1965-8 over colonial and socio-economic proposals.

The first hierarchy is a predominantly Afro-Asian scale, consisting of 56 delegations. The second hierarchy is a Latin-American scale of 19 delegations. Both hierarchies are strong scales. The Afro-Asian scale encompasses 53 delegations of the Afro-Asian caucussing group. The three other delegations were Trinidad and Tobago, Mongolia and Turkey. Only 6 of the 59 members of the Afro-Asian caucussing group were not included in this large Afro-Asian scale. Saudi Arabia (H_i = .49) and the Philippines (H_i = .40) were excluded as inclusion level for H_i was .50. Four other Afro-Asian delegations (Cambodia, Malawi, Maldive Islands and Yemen) had negative correlations with several other Afro-Asian delegations and were consequently rejected.

For the first time we found one Latin-American hierarchy that encompassed nearly the whole Latin-American group. Only three delegations were not contained in this scale, Cuba, Jamaica and Trinidad and Tobago. Jamaica and Trinidad and Tobago could have been added to the scale, had item coefficients between .50 and .30 been allowed. Mexico and to a lesser degree Chile systematically disturbed the double monotonicity in the Latin-American scale.

Thus, in the period 1965-8 two main leadership hierarchies encompassed nearly all developing nations: an Afro-Asian hierarchy including almost all Afro-Asian delegations, and a Latin-American hierarchy including almost all Latin-American delegations. Leaders of the Afro-Asian group are

TABLE 6

Cumulative scales of delegations (Stokman, 1977)

	Fraction of sponsored proposals	H_i		Fraction of sponsored proposals	H_i
Afro-Asian scale			Sierra Leone	.41	.65
Laos	.03	.79	Pakistan	.42	.62
Turkey	.05	.65	Sudan	.42	.68
Singapore	.08	.59	Kenya	.43	.73
Gambia	.08	.68	Mali	.44	.67
Burma	.09	.59	Zambia	.44	.66
Malaysia	.10	.65	Mauritania	.44	.63
Thailand	.11	.54	Iraq	.45	.63
Gabon	.12	.59	Nigeria	.45	.74
Trinidad & Tobago	.14	.54	Yugoslavia	.45	.60
Mongolia	.16	.65	Ghana	.46	.71
Central African Rep.	.19	.69	India	.47	.71
Chad	.20	.67	Algeria	.49	.74
Lebanon	.21	.58	United Arab Rep.	.50	.71
Madagascar	.24	.59	Syria	.51	.74
Iran	.24	.50	Tanzania	.52	.77
Kuwait	.25	.54	Guinea	.53	.79
Jordan	.25	.59	coefficient of scalability for the whole scale H =.63;		
Nepal	.25	.58			
Cyprus	.28	.65			
Ivory Coast	.28	.59	*Latin-American scale*		
Ceylon	.29	.51	Paraguay	.04	.91
Congo-Brazzaville	.31	.59	Brazil	.05	.61
Senegal	.31	.60	El Salvador	.05	.89
Dahomey	.31	.65	Honduras	.06	.87
Upper Volta	.31	.54	Mexico	.06	.57
Rwanda	.32	.59	Bolivia	.08	.75
Cameroon	.32	.60	Dominican Rep.	.08	.67
Afghanistan	.34	.58	Guatemala	.08	.82
Liberia	.34	.52	Haiti	.08	.74
Niger	.34	.61	Nicaragua	.08	.79
Uganda	.35	.66	Peru	.08	.76
Burundi	.35	.62	Costa Rica	.09	.64
Ethiopia	.36	.65	Panama	.09	.62
Morocco	.36	.64	Venezuela	.10	.72
Tunisia	.37	.65	Uruguay	.10	.78
Somalia	.37	.60	Argentina	.11	.74
Libya	.37	.63	Chile	.11	.60
Congo-Democratic Rep.	.38	.65	Colombia	.11	.78
Togo	.39	.63	Ecuador	.12	.79
			coefficient of scalability for the whole scale H =.74;		

Guinea, Tanzania, Syria, the UAR and Algeria; leaders of the Latin-American group are Ecuador, Colombia, Chile and Argentina.

References

General overviews can be found in: (7), (8), (11), (12), (14), (27), (31), (32), (36) and (39).
Operationalization of different types of scales: (28), (33).
Data- and measurement theory: (3), (18).
Method of paired comparisons, Thurstone's model: (4), (10), (37), (38).
Method of paired comparisons, B.T.L.-model: (22).
Likert scaling, classical test theory: (13), (21).
Guttman scaling: (5), (20), (25), (26), (35).
Mokken scaling: (17), (20), (24), (25), (27), (34).
Rasch scaling: (13), (17), (40).
Unfolding: (2), (7), (16), (19), (23).
Multidimensional scaling: (1), (9), (15), (29).
Conjoint measurement: (18).
Hierarchical models: (6), (30).

Ahrens, H.J. (1974). *Multidimensionale Skalierung*. Weinheim/Basel: Beltz Monografiën. (1)

Bechtel, G.D. (1968). 'Folded and Unfolded Scaling from Preferential Paired Comparisons', *Journal of Mathematical Psychology* 5. (2)

Bezembinder, T.G. (1970). *Van Rangorde naar Continuum*, Deventer: Van Sloghum Slaterus. (3)

Bock, R.D. and Jones, L.V. (1968). *The Measurement and Prediction of Judgment and Choice*, San Francisco: Holden-Day. (4)

Bogardus, E.S. (1933). 'A social distance scale', *Sociology and Social Research* 17. (5)

Carroll, J.D. and Pruzansky, S. (1975). Fitting of hierarchical tree structure (HTS) models, mixture of HTS models and hybrid models, via mathematical programming and alternative least squares. Presented at the US-Japan Seminar on Theory, Methods and Applications of Multidimensional Scaling and Related Techniques, San Diego. (6)

Coombs, C. (1964). *A Theory of Data*. New York: Wiley. (7)

Coombs, C., Dawes, R.M. and Tversky, A. (1970). *Mathematical Psychology, An Elementary Introduction*. Englewood Cliffs: Prentice Hall. (8)

Coxon, A.P.M. (1975). *Multidimensional Scaling*. ECPR Summer School at the University of Essex Monograph. (9)

David, F. (1954). *The Method of Paired Comparisons*. London: Griffin. (10)

Dawes, R.M. (1972). *Fundamentals of Attitude Measurement*. New York: Wiley. (11)

Edwards, A.L. (1957). *Techniques of Attitude Scale Construction*. New York: Appleton Century Crofts. (12)

Fischer, G. (1974). *Einführung in die Theorie Psychologischer Tests, Grundlagen und Anwendungen*. Bern: Huber. (13)

Fishbein, M. (ed.) (1967). *Readings in Attitude Theory and Measurement*. New York: Wiley. (14)

Green, P.E. and Carmone, F.J. (1968). *Multidimensional Scaling and Related Areas in Marketing Analysis*. Boston: Allyn and Bacon. (15)

Greenberg, M.G. (1965). 'A method of successive cumulations for the scaling of pair-comparison preference judgment', *Psychometrika* 30: 441-448. (16)

Henning, H.J. (1974). *Skalenanalyse und RASH-Modell*. Bonn Universität. (17)

Krantz, D.H., Luce, R.D., Suppes, P. and Tversky, A. (1971). *Foundations of Measurement Vol. I*. New York: Academic Press. (18)

Kruskal, J.B. (1964). 'Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis', *Psychometrika* 29: 1-27. (19)

Lazarsfield, P. and Henry, N. (1968). *Latent Structure Analysis*. New York: Houghton Mifflin.(20)
Likert, R.S. (1932). 'A technique for the measurement of attitude', *Archives of Psychology* 140.(21)
Luce, R.D. (1959). *Individual Choice Behavior*. New York: Wiley. (22)
McClelland, G.H. and Coombs, C.H. (1975). 'ORDMET: A general algorithm for constructing solutions to ordered metric structures', *Psychometrika* 40: 269-290. (23)
Mokken, R.J. (1969). 'Dutch-American comparisons of the "sense of political efficacy"', *Quality and Quantity* 3: 125-152.(24)
Mokken, R.J. (1971). *A Theory and Procedure of Scale Analysis*. Paris/The Hague: Mouton.(25)
Nesvold, B.A. (1970). 'Scalogram Analysis of Political Violence', *Comparative Political Studies* 2: 172-194.(26)
Niemöller, K. and Van Schuur, W.H. (1977). *Basic Scaling*. ECPR Summer School at the University of Essex Monograph.(27)
Robinson, J.P. et al. (1969). *Measures of Political Attitudes*. Ann Arbor: Institute for Social Research, Univ. of Michigan.(28)
Roskam, E.E.Ch.I. (1975). *Non-metric Data Analysis*. Internat. Report 75 MA 13, Psychological Laboratory, Catholic University Nijmegen.(29)
Sattah, S. and Tversky, A. (1977). 'Additive similarity trees', *Psychometrika* 42: 319-345.(30)
Scheuch, E.K. and Zehnpfennig, H. (1974). 'Skalierungserfahren in der Sozialforschung', pp.97-203 in R. König (ed.) *Handbuch der empirischen Sozialforschung band 3a: Grundlegende Methoden und Techniken, Zweiter Teil*. Stuttgart: DTV (third edition).(31)
Scott, W.A. (1968). 'Attitude Measurement', pp.204-273 in G. Lindzey and E. Aronson (eds.) *Handbook of Social Psychology Vol. II* (second edition). New York: Addison-Wesley. (32)
Shaw, M.E. and Wright, J.M. (1967). *Scales for the Measurement of Attitudes*. New York: McGraw-Hill.(33)
Stokman, Frans N. (1977). *Roll Calls and Sponsorship, A Methodological Analysis of Third World Group Formation in the United Nations*. Leyden: Sijthoff. (34)
Stouffer, S.A., et al. (1950). *Measurement and Prediction*. Princeton: Princeton University Press.(35)
Summers, G. (ed.) (1970). *Attitude Measurement*. Chicago: Rand McNally.(36)
Thurstone, L. and Chave, E.J. (1929). *The Measurement of Attitude*. Chicago: University of Chicago Press.(37)
Thurstone, L.L. (1959). *The Measurement of Values*. Chicago: Chicago University Press. (38)
Torgerson, W.S. (1958). *Theory and Methods of Scaling*. New York: Wiley.(39)
Wakenhut, R. (1974). *Messung gesellschaftlich-politischer Einstellungen*. Bern: Huber. (40)